The Role of Optics in Future High Radix Switch Design

Nathan Binkert† Al Davis† Norman P. Jouppi† Moray McLaren†
Naveen Muralimanohar† Robert Schreiber† Jung Ho Ahn‡

†HP Labs
Palo Alto, USA
{firstname.lastname}@hp.com

‡Seoul National University
Seoul, Korea
gajh@snu.ac.kr

ABSTRACT

For large-scale networks, high-radix switches reduce hop and switch count, which decreases latency and power. The ITRS projections for signal-pin count and per-pin bandwidth are nearly flat over the next decade, so increased radix in electronic switches will come at the cost of less per-port bandwidth. Silicon nanophotonic technology provides a long-term solution to this problem. We first compare the use of photonic I/O against an all-electrical, Cray YARC inspired baseline. We compare the power and performance of switches of radix 64, 100, and 144 in the 45, 32, and 22 nm technology steps. In addition with the greater off-chip bandwidth enabled by photonics, the high power of electrical components inside the switch becomes a problem beyond radix 64.

We propose an optical switch architecture that exploits high-speed optical interconnects to build a flat crossbar with multiple-writer, single-reader links. Unlike YARC, which uses small buffers at various stages, the proposed design buffers only at input and output ports. This simplifies the design and enables large buffers, capable of handling ethernet-size packets. To mitigate head-of-line blocking and maximize switch throughput, we use an arbitration scheme that allows each port to make eight requests and use two grants. The bandwidth of the optical crossbar is also doubled to to provide a $2x$ internal speedup. Since optical interconnects have high static power, we show that it is critical to balance the use of optical and electrical components to get the best energy efficiency. Overall, the adoption of photonic I/O allows 100,000 port networks to be constructed with less than one third the power of equivalent all-electronic networks. A further 50% reduction in power can be achieved by using photonics within the switch components. Our best optical design performs similarly to YARC for small packets while consuming less than half the power, and handles 80% more load for large message traffic.

Categories and Subject Descriptors:
B.4.3 [**Input/Output and Data Communications**]: Interconnections (subsystems) – *Fiber optics*;

General Terms: Throughput, Power efficiency

Keywords: Switch, Router, High-Radix, Photonics

ISCA'11, June 4–8, 2011, San Jose, California, USA.

1. INTRODUCTION

High end system performance is expected to grow by three orders of magnitude, from petascale to exascale, by 2020. The Moore's law scaling of semiconductor technology will not, by itself, meet this need; to close the gap, there will be more processing and storage components. A recent study [20] shows that an exascale system will likely have 100,000 computational nodes. The increasing scale and performance will put tremendous pressure on the network, which is rapidly becoming both a power and a performance bottleneck [21]. High-radix network switches [17] are attractive since increasing the radix reduces the number of switches required for a given system size and the number of hops a packet must travel from source to destination. Both factors contribute to reduced communication latency, component cost, and power. High-radix switches can be connected hierarchically (in topologies such as folded Clos networks [14]), directly (in a flattened butterfly or HyperX topology [2, 15]), or in a hybrid manner [16].

The chip I/O bandwidth and chip power budget are the two key limits to boosting radix. Our goal is to assess which of electronics or photonics will be better suited to overcome these limits in future switches. In order to make this assessment, we need guideposts. For electronics, we use the ITRS [27]. Since photonics has no published roadmap, we develop one as described in Section 2 and use it in a performance and power comparison between electronics and photonics.

In electronic switches, increasing radix to reduce latency while maintaining per-port bandwidth will be hard because of chip-edge bandwidth: the ITRS predicts only modest growth in per-pin bandwidth and pin count over the next decade. For example, Cray's YARC is a high-radix, high performance, single-chip switch [26], with 768 pins shared by 64 bi-directional ports, giving an aggregate bandwidth of 2.4Tb/s. Each port has three input and three output data signals, but the use of differential signaling, necessary to improve high speed signaling reliability, means that 12 pins are required in total. High speed SERDES can help by increasing the signaling rate, but this reduces the power budget available for the actual switching function. In YARC, high-speed differential SERDES consume approximately half the chip power [1].

Emerging silicon nanophotonics technology [18, 19, 22, 30, 31] may solve the pin bandwidth problem. Waveguides or fibers can be coupled directly onto on-chip waveguides, eliminating electrical data pins. While the signaling rate is comparable to that of electrical pins, high bandwidth per waveguide can be achieved with dense wavelength division multiplexing (DWDM), in which up to 64 wavelengths of light constitute independent communication channels. Because of DWDM, a high-radix photonic switch will have fewer off-chip fiber connections than pins in a comparable electronic switch. Furthermore, over a long path, an inter-

switch cable or a circuit board trace, the energy cost to send a bit of information is lower in optics than in electronics. At datacenter scale, the bit transport energy (BTE) of photonic communication is nearly independent of path length; electrical BTE grows linearly.

The next scaling limit will be power in the on-switch-chip electrical interconnect. Again, an all-electrical solution will not work. But unlike the I/O limit, the right answer is not an all-photonic solution; it is a reasonable hybrid of long-distance photonics with short-distance electronics.

On-chip global wires are increasingly slow and power hungry [12]. Global wire geometry is not scaling at the same rate as transistor geometry. To minimize fall-through latency, YARC uses repeated wires in global data and control paths. Many intermediate buffers and wires are required to support YARC's over-provisioned intra-switch bandwidth.

Photonic BTE is low, and is length independent on-chip as well as off-chip. But there are other issues. Optical modulators and receivers require constant tuning even when not being used (more in Section 2.2) resulting in static power not present in plain electrical wires. Electrical signaling over small distances can have lower BTE and be faster than optical signaling, partly due to endpoint EO/OE in optics. The distance at which optics becomes preferable will change with shrinking feature size, because electrical wires and optics scale differently. Thus a short-range electronic, long-range optical design has some justification. It should be parameterized, to adapt to the technology-dependent tradeoff.

We therefore propose a photonic architecture that employs a flat crossbar without intermediate buffers. Furthermore, we use a clustering technique, in which nearby switch ports communicate electrically over short distances to shared photonic components that connect these port clusters globally. This has the dual benefit of reducing the static-power-consuming photonic component count through sharing, and using electronic signals at short distances. To mitigate head-of-line (HOL) blocking and improve switch throughput, our arbitration scheme allows each port to make eight requests and use two grants.

Our main contributions are: 1) a photonic switch microarchitecture showcasing the importance of a careful balance of optical and electrical interconnects to maximize energy efficiency; 2) the creation of a nanophotonic roadmap; and 3) quantifying the performance and power benefits of using photonics in high-radix switch design.

2. THE ELECTRONIC AND PHOTONIC ROADMAPS

High-performance switches are not manufactured in the same volume as processors; they are relegated to older fabs. YARC, a standard cell ASIC, was fabricated in a 90 nm fab, and custom microprocessors were then fabricated in a 65 nm process [26]. Microprocessors are now fabricated in 32 nm CMOS technologies; ASICs remain at least a generation behind. We therefore focus on the 45, 32, and 22 nm CMOS technology steps.

We describe electrical and photonic I/O roadmaps. These help define the design space for high-radix switches. The electrical I/O roadmap is based on the 2009 ITRS [27]. It provides the roadmap for most switch components, but does not predict I/O power. We supplement it with SERDES power predictions based on recently published results. Although the impact of technologies such as photonics is being considered by the ITRS, there is no industry roadmap at the present time. We make a first attempt to create a photonic roadmap, based on recent literature as well as our own laboratory efforts.

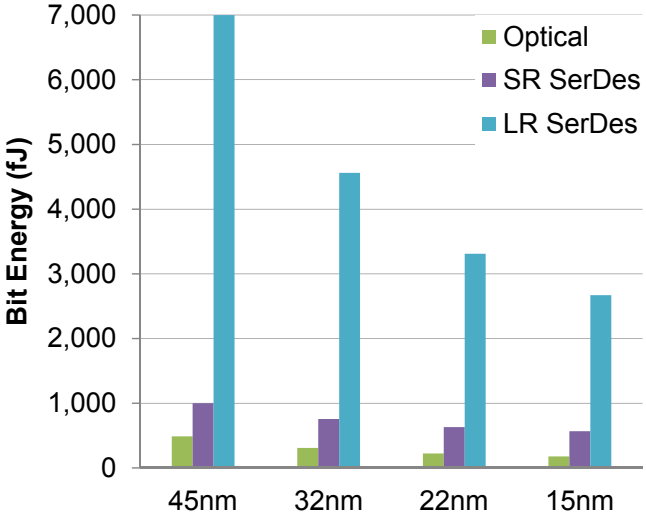


Figure 1: I/O energy per bit scaling

The benefits of photonics are compelling, but technology challenges remain before it can be deployed. Laboratory device demonstrations have been performed; waveguides, modulators, and detectors have been built and tested [7], but the ability to cheaply and reliably manufacture hundreds to millions of these devices on the same substrate has not yet been demonstrated.

2.1 Electrical I/O Roadmap

The ITRS is concerned primarily with the "short reach" or SR-SERDES, with trace lengths of a few centimeters, used for processor to main memory interconnects. Recently a number of low power SR-SERDES have been demonstrated [25, 10]. In switch applications, "long reach" or LR-SERDES are generally required so as to drive a path of up to one meter of PC board trace with at least two backplane connectors in the path. SR-SERDES use less power than LR-SERDES, but they require some form of external transceiver or buffer to drive longer transmission paths. Although switch chip power in this arrangement decreases, the overall system power grows.

Historical data show that SERDES power scales by roughly 20% per year [25]. Not all components of SERDES power will continue to scale at this rate. The external loads (impedances of off-chip wires) are not changing, and the output drive power cannot be expected to improve. Our power model for SR-SERDES and LR-SERDES takes the current state-of-the-art BTE value as its starting point. We assume that the power of the transmitter output stage remains constant, and the balance of the energy will scale according to the ITRS roadmap. The predicted BTE values for both types are shown Figure 1.

2.2 Photonic Roadmap

External transceivers cannot overcome the chip-edge bandwidth wall. An integrated technology can, by bringing light directly onto the chip. Integrated CMOS photonics, where all the components for communication with the exception of an external laser power source are integrated in a CMOS compatible process, have been demonstrated using indirect modulation [4]. However, the Mach-Zehnder modulators used in these systems are impractical for systems requiring many optical channels due to their large area and relatively high BTE.

Compact, power efficient modulators based on resonant struc-

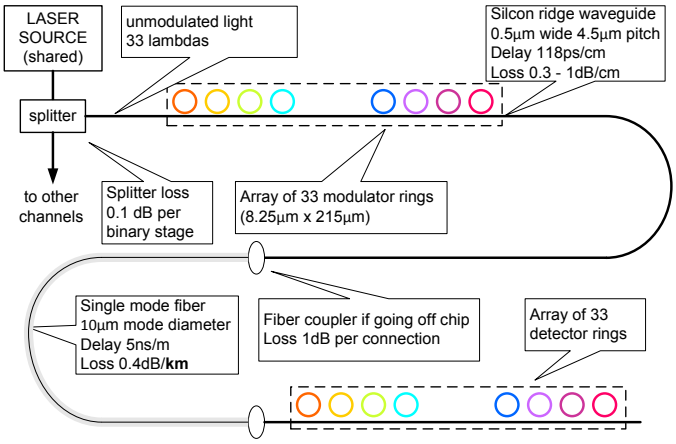


Figure 2: Interchip point-to-point DWDM link

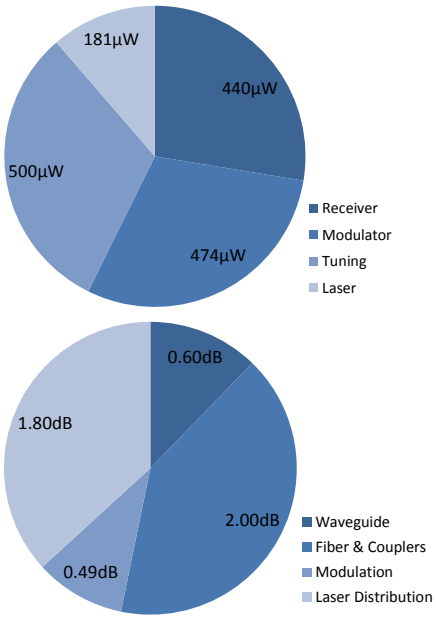


Figure 3: Point-to-point power (top) and optical losses (bottom) for 2cm of waveguide and 10m of fiber in 22nm technology

tures have been demonstrated [7]. Our proposed technology uses silicon ring resonators, similar to the devices described by Ahn et al. [3]. A ring can be used as a modulator, as a wavelength-specific switch, or as a drop filter. Rings have the additional advantage of being wavelength specific, allowing DWDM (dense wavelength division multiplexed) transmitters to be created. Rings, together with silicon ridge waveguides for on chip connectivity, waveguide-integrated germanium detectors, and grating couplers for external connectivity, constitute a complete set of components required for communications. All components can be manufactured on a common silicon substrate with the optical source being provided by an off-chip laser.

Figure 2 depicts a complete DWDM photonic link. An external mode-locked laser provides light as a "comb" of equally spaced wavelengths. An array of ring resonators in one-to-one correspondence with the wavelength comb modulates a signal on the passing light. That light is transmitted through a waveguide, into fiber via a coupler, and back into another waveguide on a different chip, and into another array of ring resonators for detection. This link can be used for both inter-chip communication via the single mode fiber or for intra-chip communication if that fiber and the related couplers are removed.

Power and losses for a complete inter-chip DWDM photonic link consisting of 2cm of waveguide and 10m of fiber are illustrated in both Figures 2 and 3. We calculate the laser electrical power from the required receiver optical input power, the total path loss including optical power distribution, and the laser efficiency. Receiver electronic power was simulated using HSPICE to model the transimpedance amplifier and limiting amplifiers. Modulation power was estimated from the measured circuit parameters of ring resonators assuming a modulation rate of 10 Gb/s at each process step. The final component is the thermal tuning power. Since all the power terms except modulation are independent of link activity, link power is not strictly proportional to usage. High speed differential electronic links exhibit a similar lack of proportionality.

3. SWITCH MICROARCHITECTURE

We chose a scalable switch microarchitecture to allow design space exploration to compare the photonic and electrical alternatives. For electrical systems, this is accomplished by increased chip pin-count and/or improved SERDES speeds, and for photonic interconnects this is enabled by the availability of more wavelengths for the WDM links. Constrained by the limits of the electrical and photonic roadmaps, we investigate switches of radix 64, 100, and 144, of each in three process generations. The decision to use a square number of ports was motivated by the desire to maintain an $N \times N$ array of subswitches in the all-electronic switch case. We view feasible designs as falling within ITRS package limitations, consuming less than 140 watts, and fitting within an 18x18 mm die. Higher power switches are possible, but would require significantly more expensive liquid conductive cooling. We view designs between 140 and 150 watts as cautionary and designs greater than 150 watts as infeasible. The die size is based on a floorplan that accounts for port interconnect pitch, input and output buffer capacity, photonic element pitch, port tile logic, and optical arbitration waveguides or electrical arbitration logic.

Datacenter switches typically conform to Ethernet style packet sizes, and vary in length from jumbo packets, commonly 9000 or more bytes, to the smallest 64 B size. For simulation purposes, we vary the packet size in multiples of 64 B, where the multiplier varies between 1 and 144. In both electronic and photonic designs, we provide buffers at both the input and output ports. Input buffers are 32, 64, and 128 KB respectively for the 45, 32, and 22 nm feature sizes. This 2x scaling tracks the 2x scaling projection of additional wavelengths. The output buffer is sized at 10 KB to accommodate an entire jumbo packet. The output buffer can also be increased in size to support link-level retry, but we are not modeling failure rates and link level retry in this work since we focus on a single switch.

For optical I/Os, we allow one input fiber and one output fiber per port, and hence the per-port bandwidths over the three process generations are 80 Gbps, 160 Gbps, and 320 Gbps. Flow control is done on a per-packet basis. The worst case inter-switch link in our model is 10 meters, and flow control must account for the round trip latency on the link plus the response time on either end. Table 1 shows the worst case number of bits that could be in fight, and the buffers are sized accordingly. Our simulations and power estimation models focus on datapath and arbitration resources. The remaining details of the various tile resources are shown in Table 1. We assume a 5 GHz electrical component clock based on ITRS [27] and drive the optical links in DDR fashion at 10 Gbps.

Table 1: Radix independent resource parameters

General	Process	nm	45	32	22
	System clock	GHz	5		
Link characteristics	Port bandwidth	Gbps	80	160	320
	Max link length	m	10		
	In flight data	Bytes	1107	2214	4428
Optical link parameters	Data wavelengths		8	16	32
	Optical data rate	Gbps	10		
Electronic link parameters	SERDES per channel bandwidth	Gbps	10	20	32
	SERDES channels per port		8	8	10
	Bit energy (LR_SERDES)	fJ/bit	7000	4560	3311
	SERDES TDP/port	mW	560	730	1060
	Electronic I/O pins/port		32	32	40
Buffers	Input buffer size	kB	32	64	64
	Header queue entries		64	128	256
	Input buffer read width	bits	32	64	128
	Input buffer write width	bits	16	32	64
	Flit size	Bytes	64		
	Packet size	Flits	1–144		
	Output buffer size	Bytes	9216		

3.1 Electronic Switch Architecture

A simple switch consists of three primary components: input buffers to store incoming messages; a crossbar to transmit the messages to the appropriate output port; and an arbiter to allocate resources and resolve conflicts. Since the latency of all three components increases with radix and size, scaling them directly to a high radix will either reduce the operating frequency or the switch throughput. Where a simple FIFO structure is used for the input buffers, a packet at the head of the buffer waiting for a busy output port will block subsequent packets from progressing even if their destination is free. This phenomenon, called head-of-line (HOL) blocking, limits the throughput of a simple crossbar switch to around 60% under uniform random traffic [13]. To address the latency problem, YARC splits crossbar traversal into three stages; 1-to-8 broadcasting (or demultiplexing) stage, 8x8 subswitch traversal stage, and 8-to-1 multiplexing stage. Buffers are inserted between stages to alleviate HOL blocking by buffering packets according to destination. A fully buffered crossbar with a dedicated buffer at every crosspoint can handle loads close to 100% of capacity. This significantly increases buffering, which grows as the square of port count. The YARC architecture reduces the buffer size requirements by partitioning the crossbar into multiple subswitches.

Figure 4 shows the organization of a distributed high-radix switch similar to YARC. The switch uses a single repeated tile with one instance for each bidirectional port. The tiles are organized as an M row by N column array, hence there are MN ports. Each tile consists of an input buffer, an N input to M output subswitch, an M input multiplexer, and an output buffer. Every subswitch has buffers at its inputs called row buffers. Every multiplexor has buffers at its inputs called column buffers. The size of these intermediate buffers is critical to avoiding HOL blocking. Packets flow from the input SERDES to the input buffer and are then sent (via a broadcast message) along the row bus to the tile that is in the same column as the output port. Note that on average, the N input buffers along a row will send one phit per cycle to each subswitch. Hence, the average load in a subswitch is only $100/N\%$. Once a phit reaches a subswitch, the first stage arbitration maps it to the tile of the correct output port. Within each column, the subswitches and output multiplexors are fully (all-to-all) connected. A second stage arbitration picks packets from the column buffers and sends them to the output buffer. This arrangement means that arbitration is local to a tile, and is limited to N inputs for the first stage and M inputs for the second stage. For electronic switch datapaths, we scale the input port bandwidth based on the roadmap we discussed in Section 2. The size of the subswitches, column, and row resources scale as the square root of the port count. For optical I/O, the output modulators and output detectors are assumed to be integrated with the tile in order to eliminate long wires and use the optical waveguides as an additional low-loss routing layer. For electronic I/O, the high speed SERDES are placed around the periphery of the chip to provide a more controlled analog environment.

3.2 Optical Switch Architecture

In the optical switch architecture, we return to a simple single level switch using an optical crossbar. This choice is motivated by the high static power of optical interconnects. YARC overprovisions wires to interconnect subswitches; they are underutilized. This is not a power efficient way of using optical interconnect.

We exploit the low propagation loss of optical waveguides to build an optical crossbar that spans the chip more power efficiently than an electronic crossbar. We address HOL blocking by using a flexible input buffer structure, and an arbitration algorithm that considers multiple requests from each input. The optical switch architecture is shown in Figure 5, with multiple I/O tiles surrounding a high-aspect-ratio optical crossbar. The I/O tile consists of a unified input buffer, output buffer, input header queue, and request generation logic.

Packets arriving on the input fiber are immediately converted into the electronic domain and stored in the input buffer. A separate header FIFO contains the routing information for every packet in the input buffer. The first eight elements of the header FIFO are visible to the request generation logic, which generates up to eight requests to the central arbiter. When a grant is received for one of the requests, the input buffer sends the relevant packet to the switch core and frees the buffer space. The input buffer has sufficient bandwidth to transfer two packets to the crossbar at a time. Since the input buffer is not FIFO, buffer space management is

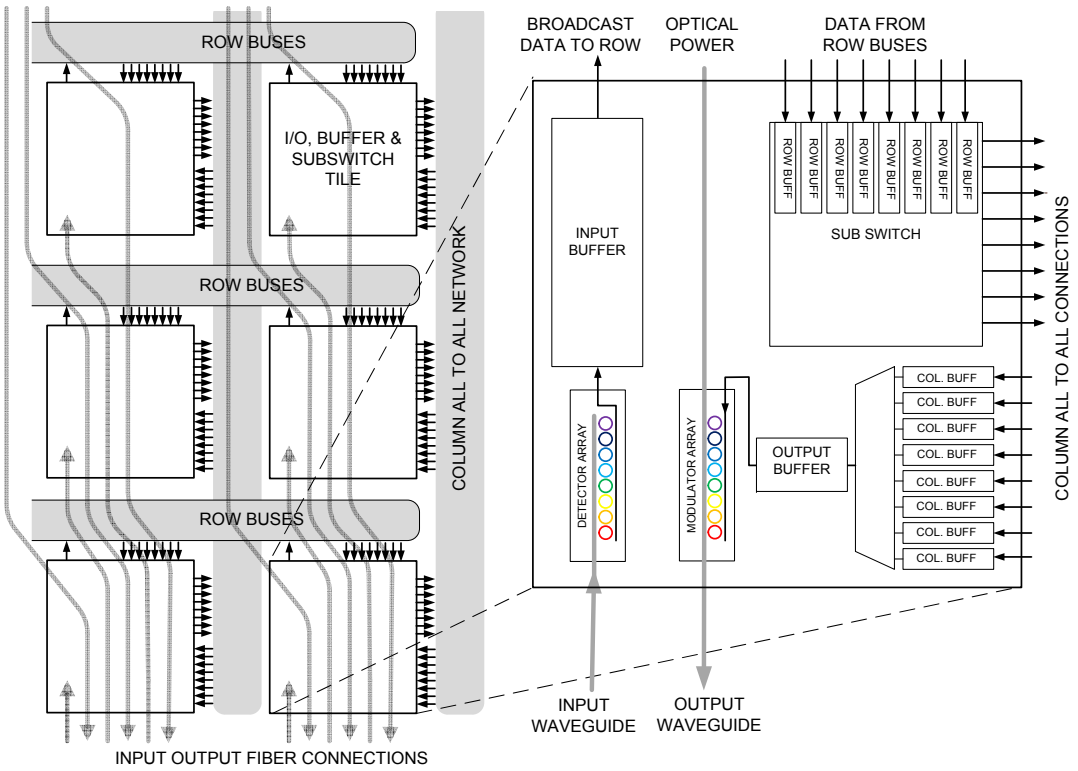


Figure 4: Array of electronic switch tiles and waveguides. Photonic I/O is incorporated into the tile.

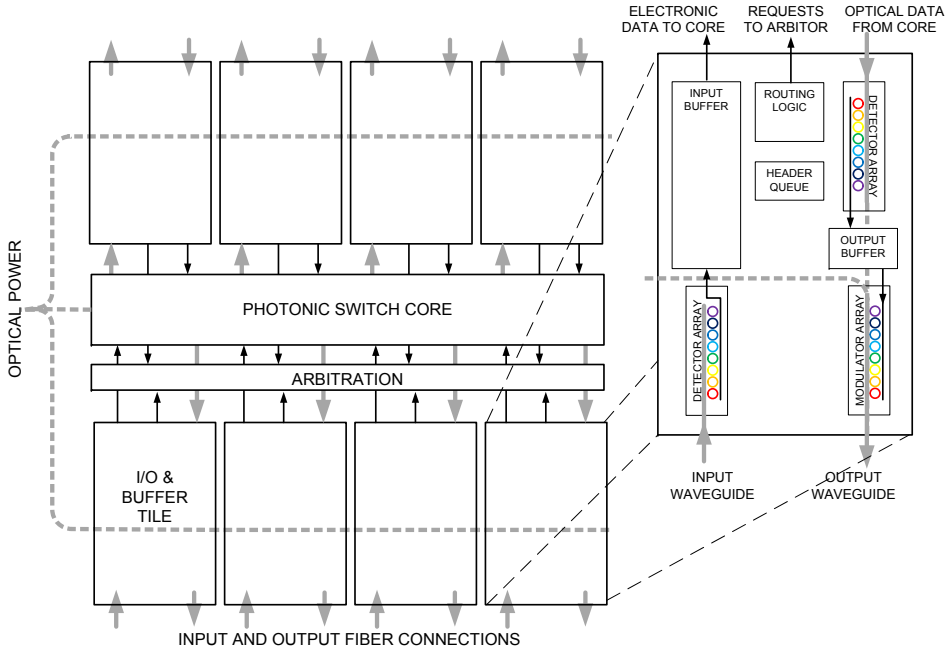


Figure 5: Tile placement for an optical switch core. A switch core with a high aspect ratio is used to exploit the low-loss of the photonic interconnect.

more complex. The crossbar operates at double the external link bandwidth, which allows the input port to "catch up" when output port contention occurs. Since the crossbar bandwidth is twice the external port bandwidth, output ports require sufficient buffering to accommodate at least one maximum-sized packet.

3.3 Optical Crossbar

A crossbar is a two-dimensional structure that broadcasts in one dimension and arbitrates in the other. In our optical crossbar, a waveguide is associated with each output port. Input port requests

Table 2: Components of optical loss

Component Name	dB
Waveguide single mode (per cm)	1
Waveguide multi mode (per cm)	0.1
Adjacent ring insertion loss	0.017
Ring scattering loss	0.001
Off-chip coupling loss	1
Non ideal beam-splitter loss [11]	0.1

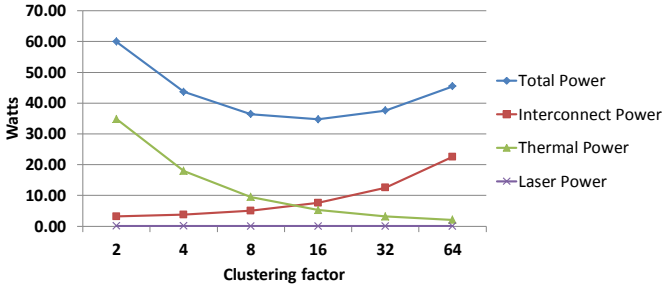


Figure 6: Varying clustering factor, radix 64 switch in 22nm technology

are granted by the arbitration structure so that at any given time only one bank of modulators will be actively driving any given waveguide. In this channel per destination approach [29], each receiver ring must always actively listen to its associated waveguide.

The tuning power for this approach scales linearly with the number of inputs, as inactive modulator arrays must be kept at a known off-frequency position to avoid interference. Multiple crossbar inputs may share a set of modulators, without impacting crossbar performance, since only one set of modulators is ever active at a time. We refer to this as *clustering*, and use this technique to minimize the number of ring resonators per waveguide.

The optical crossbar in Figure 7 shows the optical modulators shared by two pairs of inputs, one pair on each side of the optical switch, for a clustering factor of 4. Each waveguide of the 12-port switch therefore requires only three sets of modulators. We can extend the clustering factor to share the modulators between any number of adjacent tiles without impacting the throughput of the switch, but at the cost in additional electronic interconnect. The large number of rings per waveguide in the photonic crossbar means that ring related losses are more significant than for point-to-point links. Every ring induces some scattering loss, and idle, off-resonance modulator rings add loss due to adjacent partial coupling. Clustering reduces both of these loss factors. The components of loss are listed in Table 2. For the largest switch configuration studied the worst case path loss is 7.7dB.

Figure 6 shows the power savings that can be obtained by sharing the optical modulators. Initially, power drops due to the static power saved by reduced ring count. Beyond the minimum (cluster factor = 16), power grows due to the long wires in the cluster.

3.4 Thermal Tuning of Rings

A ring is resonant with a wavelength when its circumference is an integer multiple this wavelength. Manufacturing variability and thermal expansion of the silicon make it necessary to add per-ring, active temperature control to align one of the resonant frequencies of the ring with one of the wavelengths of the laser-generated comb. Watts et al. demonstrated this using Joule heating elements embedded in or near the rings [30].

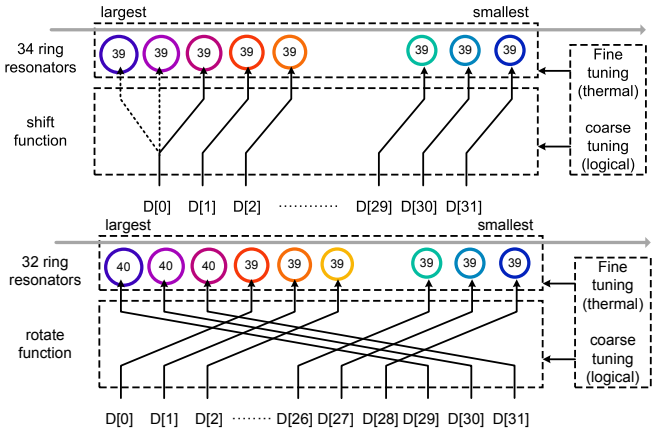


Figure 8: Coarse tuning methods to minimize heating power: (top) additional rings; and (bottom) using a higher mode.

Complete tuning flexibility for a single ring would require sufficient heating power to move the ring across a wide wavelength range. A more efficient design can minimize the thermal tuning power. One idea is to use an extended array of equally spaced rings (see the top of Figure 8). Tuning only needs to put the ring on the closest wavelength. By adding rings to extend the array, combined with a shift function between the rings and the electronic signals, the heating power required to tune between adjacent frequencies can be dramatically reduced.

A ring has multiple modes of resonance, and is said to be resonant in mode M when the effective ring path length is M times the wavelength. To avoid the added power and area costs of additional rings, with a similar reduction in maximum required heating power, we can design the geometry of the ring array such that the resonant frequency of the $M+1$th mode of the largest ring is one wavelength comb "tooth" to the low wavelength side of the shortest comb wavelength (bottom of Figure 8). In the figure, the number inside the colored ring represents the resonant mode of the ring; thus D[0] is always connected to the longest (reddest) wavelength, and D[31] to the shortest. The use of two modes in all rings gives a logical tuning range that is almost equivalent to the ring's full free spectral range, which is the frequency range between two adjacent resonant modes.

Our photonic scaling assumptions are as follows. The geometry of the rings does not scale with process improvements since ring size and resonant frequency are coupled. We assume that the modulation frequency will remain constant across generations, a consequence of the use of charge injection as the mechanism for modulation. Modulation speed in this case is limited by the carrier recombination time of the rings. A relatively low modulation rate has the advantage that simple source synchronous clocking can be used. This requires an additional clock wavelength but allows simple, low power receiver clocking when compared to high speed SERDES. We use a single added wavelength for the forwarded clock, along with groups of 8, 16, and 32 data wavelengths at the three studied process steps.

3.5 Arbitration

Our photonic crossbar design requires a high speed, low power arbiter. To better utilize the internal switch bandwidth, we performed a novel design space study using uniform random traffic to quantify the benefit that would result from increasing the number of requests and grants available for each input port. We found that

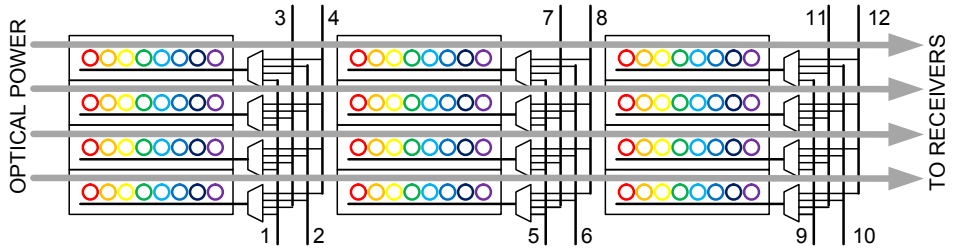


Figure 7: 12 input, 4 output crossbar with 4 way clustering

allowing 8 requests and 2 grants per port improved internal bandwidth utilization by approximately 30% on average for all radices and packet sizes. This choice allows an input port to concurrently send two packets to different output ports.

We employ two forms of electrical arbitration. The electrical arbiter (YARB) for the electrical baseline datapath is an exact replica of the distributed YARC arbitration scheme. Since our goal is to evaluate the best arbitration choice for the photonic datapath, the electrical arbiter (EARB) implementation for a photonic datapath departs from the YARC model in order to more closely mimic the optical arbitration scheme. We employ the parallel-prefix tree arbitration design of Dimitrakopoulos and Galanopoulos [9]. This approach is similar to parallel-prefix adder design, where the trick is to realize that carry-propagate and -kill are similar to a prioritized grant-propagate and -kill. The EARB contains k-tiles for each radix k configuration. Each tile is logically prioritized in a mod-k ring, where the highest priority grantee for the next selection is just after the current grantee in ring order. This provide a fairness guarantee similar to round-robin scheduling.

The EARB is centralized and pipelined, but there is little doubt that additional improvements to our current version can be found. In particular area, speed, and power improvements are likely possible with more rigorous attention to critical path timing and transistor sizing issues. Layout can be improved to reduce wire delays. Finally our current scheme uses one prefix-tree arbiter for each output port and each arbiter returns a single grant to the winning requester. Hence it is possible for an input port to receive more than two grants. When this happens, logic at the input port will select which grants are to be rejected by dropping the associated request lines. The result is that the eventual grantee will wait longer than necessary due to the extra round trip delays between the input port and arbiter.

Sending a minimum sized packet takes eight clocks. The most important aspect of any arbitration scheme is to have a round trip delay that is less than the packet transmission time. Our EARB design is optimized for delay, although we note that the dominant delay is due to the long electrical request and grant wires. Our EARB tile takes less than one 200 ps cycle for all process steps and radices. The worst case EARB request to grant time is seven clocks. The EARB power has a negligible impact on total switch power and in the worst case (radix 144, 45nm) the arbiter requires 52 pJ/operation. For 22 nm the 144 radix power is 25.7 pJ/op.

Optical arbitration uses a separate set of arbitration waveguides where a particular wavelength on an arbitration waveguide is associated with a particular egress port in the switch. We employ the *token channel* arbitration scheme proposed by Vantrease et al [29]. The optical arbitration round trip time is also less than eight clocks and the arbitration power has a negligible impact on total switch power. We conclude that there is no substantial difference between EARB and optical token channel arbitration and that either will be suitable through the 22 nm process step. Since the dominant delay component of EARB is the long request and grant wires, which grow with each new process step, we believe that in the long run optical arbitration may prove to be the winner.

Table 3: I/O and Package Constraints

	Ports	64	100	144
All optical generations	Max die size (mm)	18.1		
	Fibers per side (250μm)	72		
	Fibers per side (125μm)	144		
	Fibers required	128	200	288
	Fiber sides (250μm)	2	3	4
	Fiber sides (125μm)	1	2	2
45nm	Port Bandwidth	80Gbps		
	SERDES rate	10Gbps		
	Available SERDES pairs	600		
	Pairs Required	512	800	1152
32nm	Port Bandwidth	160Gbps		
	SERDES rate	20Gbps		
	Available SERDES pairs	625		
	Pairs Required	512	800	1152
22nm	Port Bandwidth	320Gbps		
	SERDES rate	32Gbps		
	Available SERDES pairs	750		
	Pairs Required	640	1000	1440

3.6 Packaging Constraints

We evaluated the feasibility of all the switch variants against the constraints of the ITRS roadmap for packaging and interconnect. Table 3 shows the electrical and photonic I/O resources that will be required for our choice of I/O models in all three process generations. The key conclusion is that the only feasible design for an all-electrical system capable of port bandwidths of 80 Gbps is radix 64. However even with today's 250 micron fiber packaging pitch, all of the optical I/O designs are feasible using fibers on four sides of the device. Using 125 micron pitch fiber packaging all the optical connectivity can be achieved on two sides. Even given the optimistic ITRS provisioning of high speed differential pairs, there just aren't enough to support 100 and 144 port electronic designs with the requisite port bandwidth due to packaging limitations. From a packaging perspective, the trend is clear; increasing the switch radix over the radix-64 YARC while significantly increasing bandwidth requires optical I/Os. Since power and performance are equally critical in determining feasibility, we discuss these next.

4. EXPERIMENTAL SETUP

We estimate performance with M5 [5], with new modules for the

designs we compare, modeling interactions at flit granularity. The optical model accounts for the propagation delay of light in waveguides in order to accurately quantify communication and arbitration delay.

We use CACTI 6.5 [24] to model the electronic switch and the electronic components of the photonic switch. The photonic model includes an analytic model of optical losses, input laser power, and thermal tuning power. For both, we model in detail the dominant components of the datapath, such as input and output buffers, crossbars, row and column buffers, arbiters, and vertical/horizontal buses. Other logic modules such as the Link Control Block (LCB) [26] and statistical counters contribute to the total power, but it is a negligible contribution.

In the YARC model, to calculate peak power we assume 100% load on input and output buffers. Although each subswitch can be fully loaded, the aggregate load on all the subswitches is limited by the switch's bandwidth. For example, in a switch with n subswitches handling uniform traffic, the mean load on each subswitch is no greater than $100/\sqrt{(n)}\%$, even when the switch is operating at full load. Similarly, the number of bytes transferred in horizontal and vertical buses is also limited to the aggregate I/O bandwidth.

5. RESULTS

Our initial experiments compare the performance and power of the optical full crossbar with a YARC style electronic crossbar for a range of switch sizes and traffic types. Overall, the performance results show that a YARC style electronic crossbar can perform as well as an optical crossbar, but as the radix and port bandwidth increase the power consumed by the electronic crossbar becomes prohibitive. Finally, we present power results for large networks based on the various switches that we have modeled.

5.1 Performance Results

Both switches do well on most traffic patterns, except for some contrived patterns where YARC performs poorly. Once the switch radix is large, the performance variation due to switch radix is minimal, making the performance results for all three radices roughly equivalent. The performance results also don't change appreciably at the different technology nodes. With an optical datapath, both electrical and optical arbitration schemes provide roughly the same performance because the electrical scheme is fast enough for our data points. The main benefit of the higher radix switches comes at the system level, where hop-count, switch power and cost are reduced.

Figure 9(a) shows the performance for uniform random traffic with 64 byte packets across three switch configurations at the 22nm technology node. The performance of the optical crossbar with and without speedup brackets the YARC design. The optical crossbar, without speedup, is performance limited by its inability to catch up when an input is unable to send to a particular output due to contention. Though YARC also doesn't have internal speedup, the column wires, being independent resources, in effect give the output half of the switch significant speedup. With very large input buffers, the YARC design is easily able to keep its row buffers filled and thus output contention never propagates back to the input stage. The increase in latency with the applied load is almost identical for both approaches reflecting the fact that although the YARC is a multistage design, the use of minimal shared internal resources means that it performs as well as a full crossbar.

Figure 9(b) shows the performance for jumbo packets. With jumbo packets, there are two problems with the YARC design which prevent high throughput. First, the row buffers are too small to store an entire packet, so congestion at the output causes the packet to trail back through the switch and the row bus and results in HOL blocking. Since we are targeting switches for Ethernet networks, flits cannot be interleaved because packets must be single units. We can fix this HOL blocking by providing credit-based flow control from input to output, but even with zero-latency flow control this doesn't improve the load that the switch can handle because the switch is unable to keep the column buffers full, thus losing its ability to catch up when there is output contention. The optical crossbar without internal speedup does better with large packets because the duration of output contention is short compared to the duration of packet transmission (i.e. a failed arbitration might cause a few cycle loss of bandwidth whereas the data transmission takes hundreds of cycles).

Generation	Port BW	Core Type	Radix 64	Radix 100	Radix 144
45nm	80Gbps	Electronic	41.8	72.7	120.7
		Optical	13.2	17.4	31.9
32nm	160Gbps	Electronic	38.0	65.9	109.0
		Optical	22.9	27.7	50.9
22nm	320Gbps	Electronic	52.4	91.9	153.8
		Optical	34.2	41.3	76.3

Table 4: Switch core power in watts

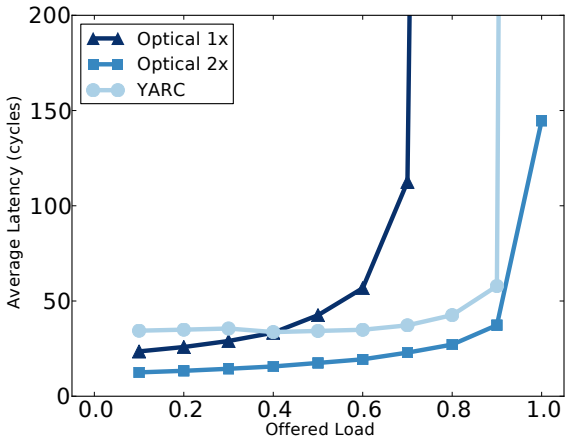
Generation	Port BW	Switch Core	I/O	Radix 64	Radix 100	Radix 144
45nm	80Gbps	E	E	77.6	128.7	201.4
		E	O	44.1	76.3	125.9
		O	O	15.5	21.0	37.0
32nm	160Gbps	E	E	89.7	146.7	225.3
		E	O	40.9	70.4	115.5
		O	O	25.8	32.2	57.5
22nm	320Gbps	E	E	135.3	221.5	340.4
		E	O	56.3	98.0	162.6
		O	O	38.1	47.4	85.1

Table 5: Overall switch power including I/O in watts

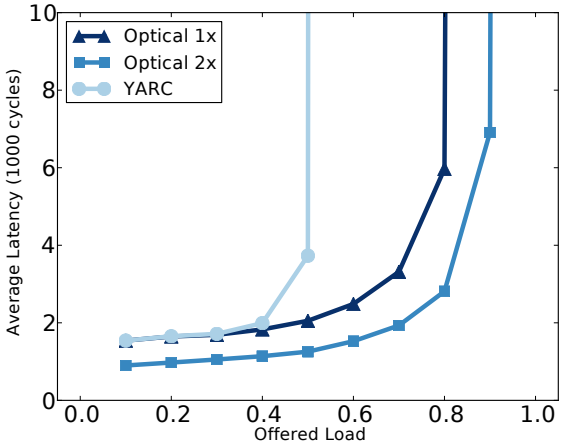
5.2 Power Results

Table 4 compares the peak power for optical and electronic switch cores for various switch sizes and technology generations. It is clear that across all technology nodes optical cores consume less power. In many cases the electrical switch power is very high, so that even if we break the pin barrier with optical off-chip interconnects, it is not feasible to build high-bandwidth, high-radix electric switches without incurring exorbitant cooling costs.

Compared to electrical switch cores, optical core power increases more slowly with radix. In electrical switches, the buffered crossbar design is a key to enabling high throughput. But its complexity grows quadratically with radix, leading to high power consumption. The row/column interconnects, consisting of fast repeated wires switching at high frequency, contribute heavily to the total power in electrical switches. Optical switch cores overcome both these problems by leveraging superior characteristics of optical interconnect and our novel arbitration scheme. The proposed 8-request, 2-grant scheme is able to achieve high throughput without intermediate buffers. The optical crossbar is effective in reducing the communication overhead. The only optical component that scales nonlinearly is the laser power (due to the loss in the link), but its contribution to the total power is minimal. The clustering technique helps keep the laser power contribution low even for high radices by reducing the number of optical rings required.



(a) Uniform random traffic, 64 byte packets, and 22nm technology. The 1x and 2x refer to the internal speedup of the optical switch.



(b) Uniform random traffic, 9216 byte packets, and 22nm technology. The 1x and 2x refer to the internal speedup of the optical switch.

Figure 9: Switch throughput comparison

Table 5 shows the total power including I/O for all configurations. For high port count, devices with electronic I/O become impractical. Across the design space, electronic switch cores are considered feasible if the total power consumed is within 140W. Beyond this threshold, more expensive conductive liquid cooling is required. Hence for high port count designs, the optical switch core has a considerable power advantage. Packaging requirements make the case even stronger for photonics.

Figure 10 shows the per-bit energy for large scale HyperX networks [2] for a range of switch components in the 22nm generation. This shows a double advantage of photonic I/O in both reducing power and enabling higher radix switches; switches of greater than 64 ports with electronic I/O exceed practical device power limits and packaging constraints. The combination of greater radix and lower component power leads to a factor-of-three savings in interconnect power for large networks using photonic I/O. A further 2x power savings can be realized by exploiting photonics for the switch core. When photonics is applied in our channel per destination approach, the tuning power of idle modulator rings becomes the most significant power overhead.

6. RELATED WORK

Single-chip CMOS high-radix Ethernet switches with up to 64 ports have recently become available [6, 8]. A significant fraction of the silicon area and power consumption in these devices is associated with the complexity of Ethernet routing. In this work we assume a simplified, compact addressing scheme, to avoid the need for content addressable memories for routing tables in a sparse address space. In a multistage network used for Ethernet traffic, the function of translating between standards-based addressing schemes and the compact scheme is required only at the ingress side of the network. This saves power on inter-switch transfers and enables larger switches to be constructed due to the lower routing overhead.

Recent work has studied the design challenges of building high-radix single chip switches. Mora et al. [23] propose partitioning the crossbar in half to improve scalability. We follow Kim et al. [17] by using a deeper hierarchical structure to construct electronic switch cores. A more detailed discussion on the implementation of the YARC switch is contained in [26].

The state of the art for CMOS integrated photonics today is limited to simple transceiver devices [4]. Krishnamoorthy, et al. [21] demonstrate some of the component technologies for larger scale integrated CMOS photonics in chip-to-chip applications. However this work is focused on the use of photonics to build photonically enabled macrochips, rather than components for use in data center networks. The use of integrated photonics for intra-chip communication is the subject of much current research. Shacham, et al. [28] propose an on-chip optical network for core-to-core communication. In this case, the switching function is optical circuit switching with an optical path being established between the communicating cores. While this can be more power efficient for long transfers, it is less efficient for heavy short-packet loads.

7. CONCLUSIONS

Integrated CMOS photonic I/O permits the scaling of switch radix beyond the electrical pin and power limitations of projected CMOS technology. Once we break the pin barrier and scale beyond 64 high-bandwidth ports, on-chip global wires create a serious power problem. To address this we propose a novel optical switch architecture that uses a flat optical crossbar. We show that by leveraging high bandwidth optical waveguides to provide significant internal speedup, and by using an arbitration scheme that takes eight requests and grants two, we overcome HOL blocking. To reduce the high static power of optics, we share photonic components in a way that balances the use of optics and electrical wires. Our architecture restricts the use of buffers to just input and output ports, and this makes it feasible to size them adequately to handle jumbo packets common in Ethernet switches. A detailed analysis shows that our proposals can reduce the system power in several ways: 1) by adopting optical I/O, we can reduce the switch power by up to 52%; 2) by using optical interconnects on-chip, we can get another 47% reduction in power at 22nm and radix 144, bringing the overall chip power well below 150W; and 3) by clustering rings and sharing them among ports, we can reduce the switch power by 41% in a radix 64 switch. Photonics, due to its low power and the reduced component count that high radix switches provide, can improve by a factor of six the energy per bit of a 100,000 port interconnection network when compared to an all electrical implementation.

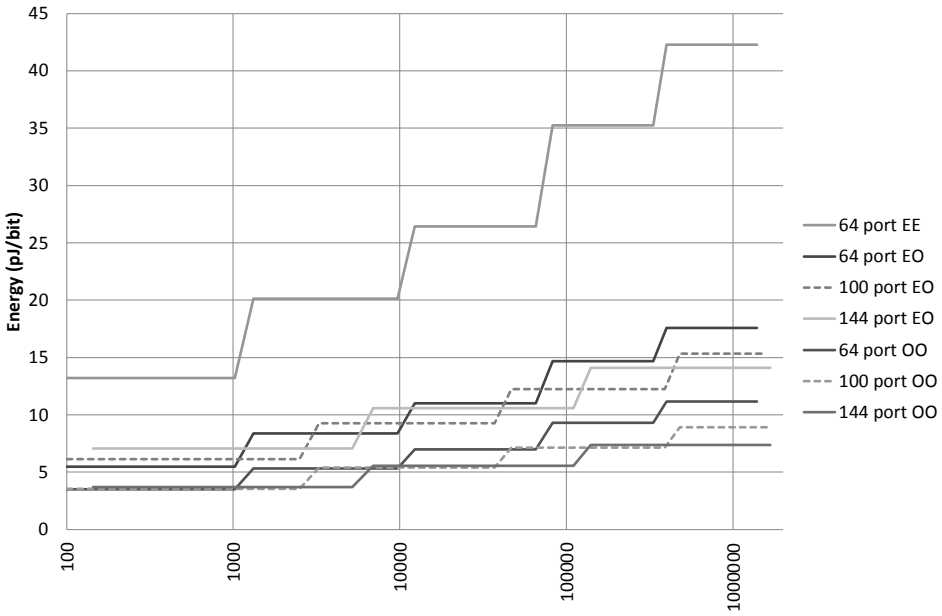


Figure 10: Energy per bit of HyperX networks with various numbers of terminals. (22nm, 320Gbps ports)

8. ACKNOWLEDGMENTS

Jung Ho Ahn was supported in parts by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0003683).

9. REFERENCES

[1] 2010. Mike Parker, personal communication.

[2] J. Ahn, N. Binkert, A. Davis, M. McLaren, and R. S. Schreiber. HyperX: Topology, Routing, and Packaging of Efficient Large-Scale Networks. *Supercomputing*, Nov. 2009.

[3] J. Ahn, M. Fiorentino, R. Beausoleil, N. Binkert, A. Davis, D. Fattal, N. Jouppi, M. McLaren, C. Santori, R. Schreiber, S. Spillane, D. Vantrease, and Q. Xu. Devices and architectures for photonic chip-scale integration. *Applied Physics A: Materials Science & Processing*, 95(4):989–997, 2009.

[4] B. Analui, D. Guckenberger, D. Kucharski, and A. Narasimha. A Fully Integrated 20-Gb/s Optoelectronic Transceiver Implemented in a Standard 0.13 micron CMOS SOI Technology. *IEEE Journal of Solid-State Circuits*, 41(25):2945–2955, Dec 2006.

[5] N. L. Binkert, R. G. Dreslinski, L. R. Hsu, K. T. Lim, A. G. Saidi, and S. K. Reinhardt. The M5 Simulator: Modeling Networked Systems. *IEEE Micro*, 26(4):52–60, Jul/Aug 2006.

[6] Broadcom. BCM56840 Series High Capacity StrataXGS®Ethernet Switch Series. http://www.broadcom.com/products/Switching/Data-Center/BCM56840-Series.

[7] L. Chen, K. Preston, S. Manipatruni, and M. Lipson. Integrated GHz silicon photonic interconnect with micrometer-scale modulators and detectors. *Optical Express*, 17(17):15248–15256, 2009.

[8] U. Cummings. FocalPoint: A Low-Latency, High-Bandwidth Ethernet Switch Chip. In *Hot Chips 18*, Aug 2006.

[9] G. Dimitrakopoulos and K. Galanopoulos. Fast Arbiters for On-Chip Network Switches. In *International Conference on Computer Design*, pages 664–670, Oct 2008.

[10] K. Fukuda, H. Yamashita, G. Ono, R. Nemoto, E. Suzuki, T. Takemoto, F. Yuki, and T. Saito. A 12.3mW 12.5Gb/s complete transceiver in 65nm CMOS. In *ISSCC*, pages 368–369, Feb 2010.

[11] S. J. Hewlett, J. D. Love, and V. V. Steblina. Analysis and design of highly broad-band, planar evanescent couplers. *Optical and Quantum Electronics*, 28:71–81, 1996. 10.1007/BF00578552.

[12] R. Ho. *On-Chip Wires: Scaling and Efficiency*. PhD thesis, Stanford University, August 2003.

[13] M. Karol, M. Hluchyj, and S. Morgan. Input versus output queueing on a space-division packet switch. *Communications, IEEE Transactions on*, 35(12):1347 – 1356, Dec. 1987.

[14] J. Kim, W. J. Dally, and D. Abts. Adaptive Routing in High-Radix Clos Network. In *SC'06*, Nov 2006.

[15] J. Kim, W. J. Dally, and D. Abts. Flattened Butterfly: a Cost-efficient Topology for High-Radix Networks. In *ISCA*, Jun 2007.

[16] J. Kim, W. J. Dally, S. Scott, and D. Abts. Technology-Driven, Highly-Scalable Dragonfly Topology. In *ISCA*, Jun 2008.

[17] J. Kim, W. J. Dally, B. Towles, and A. K. Gupta. Microarchitecture of a High-Radix Router. In *ISCA*, Jun 2005.

[18] N. Kirman, M. Kirman, R. K. Dokania, J. F. Martinez, A. B. Apsel, M. A. Watkins, and D. H. Albonesi. Leveraging Optical Technology in Future Bus-based Chip Multiprocessors. In *MICRO*, pages 492–503, 2006.

[19] B. R. Koch, A. W. Fang, O. Cohen, and J. E. Bowers. Mode-locked silicon evanescent lasers. *Optics Express*, 15(18):11225, Sep 2007.

[20] P. M. Kogge (editor). Exascale computing study: Technology challenges in achieving exascale systems. Technical Report TR-2008-13, University of Notre Dame, 2008.

[21] A. Krishnamoorthy, R. Ho, X. Zheng, H. Schwetman, J. Lexau, P. Koka, G. Li, I. Shubin, and J. Cunningham. The integration of silicon photonics and vlsi electronics for computing systems. In *Photonics in Switching, 2009. PS '09. International Conference on*, pages 1 –4, 2009.

[22] M. Lipson. Guiding, Modulating, and Emitting Light on Silicon–Challenges and Opportunities. *Journal of Lightwave Technology*, 23(12):4222–4238, Dec 2005.

[23] G. Mora, J. Flich, J. Duato, P. López, E. Baydal, and O. Lysne. Towards an efficient switch architecture for high-radix switches. In *Proceedings of the 2006 ACM/IEEE symposium on Architecture for Networking and Communications Systems*, pages 11–20, 2006.

[24] N. Muralimanohar, R. Balasubramonian, and N. Jouppi. Optimizing NUCA Organizations and Wiring Alternatives for Large Caches with CACTI 6.0. In *MICRO*, Dec 2007.

[25] R. Palmer, J. Poulton, W. J. Dally, J. Eyles, A. M. Fuller, T. Greer, M. Horowitz, M. Kellam, F. Quan, and F. Zarkeshvarl. A 14mW 6.25Gb/s Transceiver in 90nm CMOS for Serial Chip-to-Chip Communications. In *ISSCC*, Feb 2007.

[26] S. Scott, D. Abts, J. Kim, and W. J. Dally. The Black Widow High-Radix Clos Network. In *ISCA*, Jun 2006.

[27] Semiconductor Industries Association. International Technology Roadmap for Semiconductors. http://www.itrs.net, 2009 Edition.

[28] A. Shacham, K. Bergman, and L. P. Carloni. On the Design of a Photonic Network-on-Chip. In *NOCS*, pages 53–64, 2007.

[29] D. Vantrease, N. Binkert, R. S. Schreiber, and M. H. Lipasti. Light Speed Arbitration and Flow Control for Nanophotonic Interconnects. In *MICRO*, Dec 2009.

[30] M. R. Watts, W. A. Zortman, D. C. Trotter, G. N. Nielson, D. L. Luck, and R. W. Young. Adiabatic Resonant Microrings (ARMs) with Directly Integrated Thermal Microphotonics. 2009.

[31] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson. Micrometre-Scale Silicon Electro-Optic Modulator. *Nature*, 435:325–327, May 2005.